Unit 7: Multiple linear regression

2. Model Selection, Diagnostics, Transformations & Case Study

Sta 101 – Spring 2019

Duke University, Department of Statistical Science

Dr. Ellison

Slides posted at
https://www2.stat.duke.edu/courses/Spring19/sta101.001/

Outline

1. Housekeeping
2. Application Exercise 7.1
3. Transformations
4. Case Study

Coming up…

- Project Stage 2 Materials due 4/17 11:55pm (html file, RMD file, slides)
- Project Stage 2 Presentations 4/18
- Problem Set 7 and Performance Assessment 7 Sunday 4/21 11:55pm

---

- Don't forget to **ask/answer 2 questions on Piazza** before the final exam… part of your participation grade! Memes don't count ☺
- TA office hours officially end after 4/24!
- 1-2 review sessions outside of class during reading period… scheduling poll coming soon!

Outline

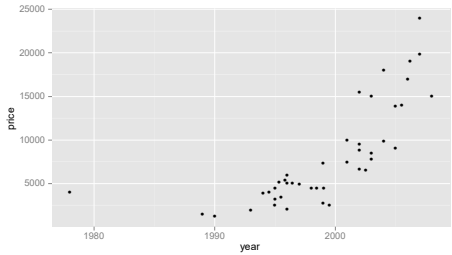
1. Housekeeping
2. Application Exercise 7.1
3. Transformations
4. Case Study

Outline

1. Housekeeping
2. Application Exercise 7.1
3. Transformations
4. Case Study

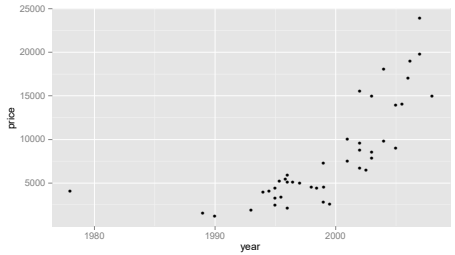
Outline

What can we *try* if our linear model does not meet all the conditions?



Truck prices

The scatterplot below shows the relationship between year and price of a random sample of 43 pickup trucks. Describe the relationship between these two variables.



From: http://faculty.chicagobooth.edu/robert.gramacy/teaching.html

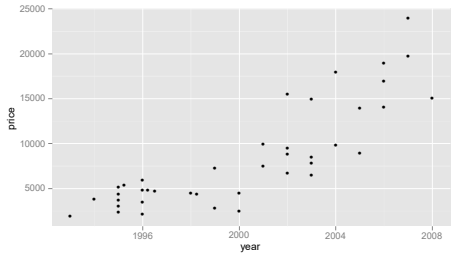
Remove unusual observations

Let's remove trucks older than 20 years, and only focus on trucks made in 1992 or later.

Remove unusual observations

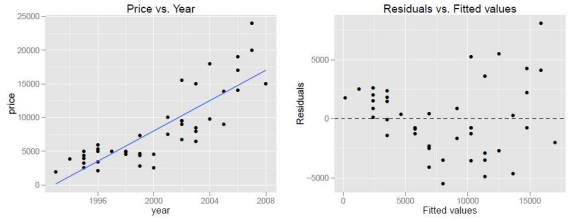
Let's remove trucks older than 20 years, and only focus on trucks made in 1992 or later.

Now what can you say about the relationship?



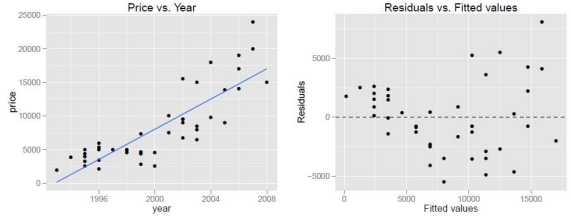
Truck prices - linear model?

Model: $\widehat{price} = b_0 + b_1\ year$



Truck prices - linear model?

Model: $\widehat{price} = b_0 + b_1\ year$

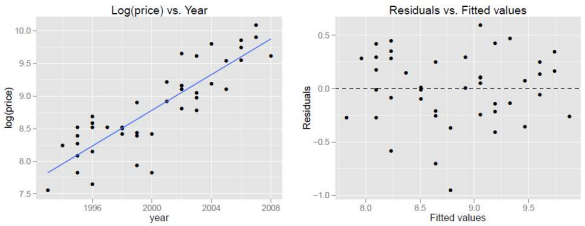


The linear model doesn't appear to be a good fit since the residuals have non-constant variance.

Truck prices - log transform of the response variable

Model: $\widehat{log(price)} = b_0 + b_1\ year$

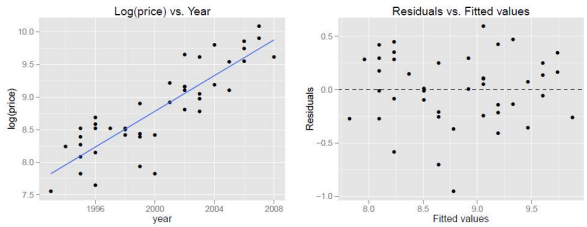
*In this class and many other statistics classes, we refer to "log( )" as the natural log (ie: ln( ) or $\log_e$( ).)



Truck prices - log transform of the response variable

Model: $\widehat{log(price)} = b_0 + b_1\ year$

*In this class and many other statistics classes, we refer to "log( )" as the natural log (ie: ln( ) or $\log_e$( ).)



We applied a log transformation to the response variable. The relationship now seems linear, and the residuals no longer have non-constant variance.

Interpreting models with log transformation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-265.073	25.042	-10.585	0.000
year	0.137	0.013	10.937	0.000

Interpreting models with log transformation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-265.073	25.042	-10.585	0.000
year	0.137	0.013	10.937	0.000

Model: $\widehat{\log(price)} = -265.073 + 0.137\ year$

Interpreting models with log transformation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-265.073	25.042	-10.585	0.000
year	0.137	0.013	10.937	0.000

Model: $\widehat{\log(price)} = -265.073 + 0.137\ year$

- **Less Useful Interpretation (using the standard way of interpreting regression slopes):** "For each additional year the car is newer (for each year decrease in car's age) we would expect the log price of the car to increase on average by 0.137 log dollars."
- which is not very useful... ?

Working with logs

- Subtraction and logs: $log(a) - log(b) =$

*In this class and many other statistics classes, we refer to "log( )" as the natural log (ie: ln( ) or $\log_e$( ).)

Working with logs

- Subtraction and logs: $log(a) - log(b) = log(\frac{a}{b})$

*In this class and many other statistics classes, we refer to "log( )" as the natural log (ie: ln( ) or $\log_e$( ).)

Working with logs

- Subtraction and logs: $log(a) - log(b) = log(\frac{a}{b})$
- Natural logarithm: $e^{log(x)} =$

*In this class and many other statistics classes, we refer to "log( )" as the natural log (ie: ln( ) or $\log_e$( ).)

Working with logs

- Subtraction and logs: $log(a) - log(b) = log(\frac{a}{b})$
- Natural logarithm: $e^{log(x)} = x$

*In this class and many other statistics classes, we refer to "log( )" as the natural log (ie: ln( ) or $\log_e$( ).)

Working with logs

- Subtraction and logs: $log(a) - log(b) = log(\frac{a}{b})$
- Natural logarithm: $e^{log(x)} = x$
- We can these identities to "undo" the log transformation **for better interpretation.**

*In this class and many other statistics classes, we refer to "log( )" as the natural log (ie: ln( ) or $\log_e$( ).)

Interpreting models with log transformation (cont.)

The Less Useful Interpretation "For each additional year the car is newer (for each year decrease in car's age) we would expect the log price of the car to increase on average by 0.137 log dollars."

Putting this in mathematical equation form:

$$\log(\text{price at year } x + 1) - \log(\text{price at year } x) = 0.137$$

Interpreting models with log transformation (cont.)

Manipulating this equation *(with log properties)* to get a better interpretation of a slope:

$$\log(\text{price at year } x+1) - \log(\text{price at year } x) = 0.137$$

Interpreting models with log transformation (cont.)

Manipulating this equation *(with log properties)* to get a better interpretation of a slope:

$$\log(\text{price at year } x+1) - \log(\text{price at year } x) = 0.137$$

$$log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right) = 0.137$$

Logarithm properties

Interpreting models with log transformation (cont.)

Manipulating this equation *(with log properties)* to get a better interpretation of a slope:

$$\log(\text{price at year } x+1) - \log(\text{price at year } x) = 0.137$$

$$\log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right) = 0.137$$

$$e^{\log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right)} = e^{0.137}$$

Logarithm properties

Interpreting models with log transformation (cont.)

Manipulating this equation *(with log properties)* to get a better interpretation of a slope:

$$\log(\text{price at year } x+1) - \log(\text{price at year } x) = 0.137$$

$$\log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right) = 0.137$$

$$e^{\log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right)} = e^{0.137}$$

$$\frac{\text{price at year } x+1}{\text{price at year } x} = 1.15$$

Logarithm properties

Interpreting models with log transformation (cont.)

Manipulating this equation *(with log properties)* to get a better interpretation of a slope:

$$\log(\text{price at year } x+1) - \log(\text{price at year } x) = 0.137$$

$$\log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right) = 0.137$$

$$e^{\log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right)} = e^{0.137}$$

$$\frac{\text{price at year } x+1}{\text{price at year } x} = 1.15$$

How do we interpret this now?

Interpreting models with log transformation (cont.)

Manipulating this equation *(with log properties)* to get a better interpretation of a slope:

$$\log(\text{price at year } x+1) - \log(\text{price at year } x) = 0.137$$

$$\log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right) = 0.137$$

$$e^{\log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right)} = e^{0.137}$$

$$\frac{\text{price at year } x+1}{\text{price at year } x} = 1.15$$

Correct Interpretation: "For each additional year the car is newer (for each year decrease in car's age) we would expect the *price of the car*, on average, to increase *by a factor of 1.15*."

Interpreting models with log transformation (cont.)

Manipulating this equation *(with log properties)* to get a better interpretation of a slope:

$$\begin{aligned}
\log(\text{price at year } x+1) - \log(\text{price at year } x) &= 0.137 \\
log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right) &= 0.137 \\
e^{log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right)} &= e^{0.137} \\
\frac{\text{price at year } x+1}{\text{price at year } x} &= 1.15
\end{aligned}$$

Correct Interpretation: "For each additional year the car is newer (for each year decrease in car's age) we would expect the *price of the car*, on average, to increase *by 15%*."

Interpreting models with log transformation (cont.)

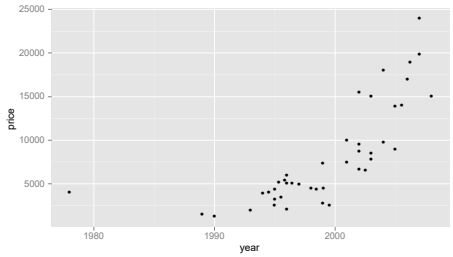
Manipulating this equation *(with log properties)* to get a better interpretation of a slope:

$$\begin{aligned}
\log(\text{price at year } x+1) - \log(\text{price at year } x) &= 0.137 \\
log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right) &= 0.137 \\
e^{log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right)} &= e^{0.137} \\
\frac{\text{price at year } x+1}{\text{price at year } x} &= 1.15
\end{aligned}$$

~~**INCORRECT Interpretation:** For each additional year the car is newer (for each year decrease in car's age) we would expect the price of the car, on average, to increase on average by *1.15 dollars*.~~

Outline

What can we *try* if our linear model does not meet all the conditions?



Recap: dealing with non-constant variance

- **Non-constant variance** is one of the most common model violations, however it is **usually fixable by transforming the response (y) variable**

Recap: dealing with non-constant variance

- **Non-constant variance** is one of the most common model violations, however it is **usually fixable by transforming the response (y) variable**
- The most common variance stabilizing transform is the log transformation: $log(y)$, especially *useful when the response variable is (extremely) right skewed*.

Recap: dealing with non-constant variance

- When using a log transformation on the response variable the interpretation of the slope changes:
  - Numerical Explanatory Variable x Interpretation:
    - $e^{b1}<1$
      - "For each unit increase in x, y is expected, on average, to decrease by a factor of e^{b1}."
      - "For each unit increase in x, y is expected, on average, to decrease $(1-e^{b1})$%".
    - $e^{b1}>1$
      - "For each unit increase in x, y is expected, on average, to increase by a factor of e^{b1}."
      - "For each unit increase in x, y is expected, on average, to increase $(e^{b1}-1)$%".

Recap: dealing with non-constant variance

- When using a log transformation on the response variable the interpretation of the slope changes:
  - Categorical Predictor Interpretation:
    - $e^{b1}<1$
      - "This predictor level is expected to have, on average, lower y than the reference level by a factor of e^{b1}."
      - "This predictor level is expected to have, on average, a $(1-e^{b1})$% lower y than the reference level.
    - $e^{b1}>1$
      - "This predictor level is expected to have, on average, higher y than the reference level by a factor of e^{b1}."
      - "This predictor level is expected to have, on average, a $(e^{b1}-1)$% higher y than the reference level.

Recap: dealing with non-constant variance

- Another useful transformation is the square root: $\sqrt{y}$ especially useful when the **response variable is counts**.
- These transformations may also be useful when the relationship is non-linear, but in those cases *a polynomial regression may also be needed*.

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

Outline

1. Housekeeping
2. Application Exercise 7.1
3. Transformations
4. Case Study

Data from the ACS

1. `income`: Yearly income (wages and salaries)
2. `employment`: Employment status, not in labor force, unemployed, or employed
3. `hrs_work`: Weekly hours worked
4. `race`: Race, White, Black, Asian, or other
5. `age`: Age
6. `gender`: gender, male or female
7. `citizens`: Whether respondent is a US citizen or not
8. `time_to_work`: Travel time to work
9. `lang`: Language spoken at home, English or other
10. `married`: Whether respondent is married or not
11. `edu`: Education level, hs or lower, college, or grad
12. `disability`: Whether respondent is disabled or not
13. `birth_qrtr`: Quarter in which respondent is born, jan thru mar, apr thru jun, jul thru sep, or oct thru dec

Load and subset data

```
acs_emp <- acs %>%
  filter(employment == "employed", income > 0)
```

Aside: categorical (factor) variables in R

```
acs_emp %>%
  select(employment) %>%
  table()
```

Aside: categorical (factor) variables in R

```
acs_emp %>%
  select(employment) %>%
  table()
```

```
not in labor force       unemployed         employed
                 0                0              787
```

Aside: categorical (factor) variables in R

```
acs_emp %>%
  select(employment) %>%
  table()
```

```
not in labor force       unemployed         employed
                 0                0              787
```

```
acs_emp <- droplevels(acs_emp) # overwrite acs_emp

acs_emp %>%
  select(employment) %>%
  table()
```

The inference() function may give you errors if your dataframe has had some of it's original levels filtered completely out and you don't run the **droplevels() function.*

```
employed
     787
```

Full model

Suppose we only want to consider the following explanatory variables: `hrs_work`, `race`, `age`, `gender`, `citizen`.

```
m_full = lm(income ~ hrs_work + race + age + gender
    + citizen, data = acs_emp)
```

Full model

Suppose we only want to consider the following explanatory variables: `hrs_work`, `race`, `age`, `gender`, `citizen`.

```
m_full = lm(income ~ hrs_work + race + age + gender
    + citizen, data = acs_emp)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-17215.60	11399.81	-1.51	0.13
hrs_work	1251.31	153.14	8.17	0.00
raceblack	-13202.39	6373.05	-2.07	0.04
raceasian	32699.34	8903.66	3.67	0.00
raceother	-12032.88	7556.78	-1.59	0.11
age	760.99	129.71	5.87	0.00
genderfemale	-17246.91	3887.17	-4.44	0.00
citizenyes	-9537.20	8360.85	-1.14	0.25

Diagnostics -- code

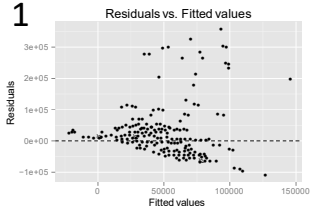
```
m_full = lm(income ~ hrs_work + race + age + gender
    + citizen, data = acs_emp)
```

```
# residuals vs. fitted
qplot(data = m_full, y = .resid, x = .fitted, geom = "point") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals") +
  ggtitle("Residuals vs. Fitted values")

# histogram of residuals
qplot(data = m_full, x = .resid, geom = "histogram") +
  xlab("Residuals") +
  ggtitle("Histogram of residuals")

# normal prob plot of residuals
qplot(data = m_full, sample = .resid, stat = "qq") +
  ggtitle("Normal probability plot of residuals")

# order of residuals
qplot(data = m_full, y = .resid) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  ylab("Residuals") +
  xlab("Order of data collection") +
  ggtitle("Residuals vs. Order of data collection")
```

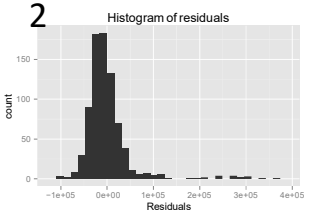
Diagnostics

What do you think? *(Response Variable = **Income**)*

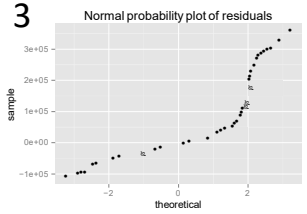
1



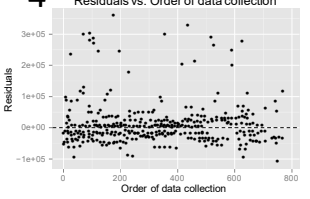
2



3



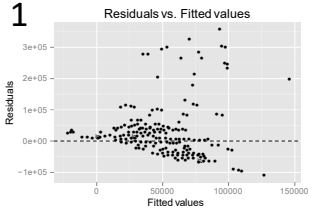
4



Diagnostics

What do you think? *(Response Variable = **Income**)*

1

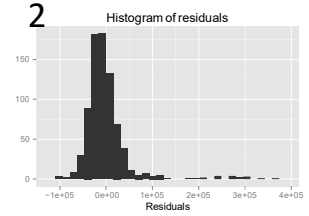


- **Linearity Condition**: Scatter mostly above the line, then mostly below the line as fitted values increases → **most likely not linear**.
- **Constant Variance of Residuals Condition**: Variance of residuals increases (not constant) as fitted values increases → **constant variance of residual condition not met.**

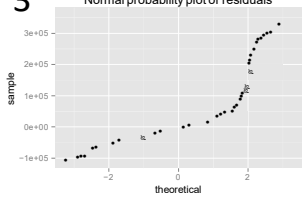
Diagnostics

What do you think? *(Response Variable = **Income**)*

2



3



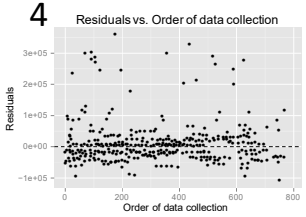
- **Normality of Residuals Condition**: The distribution of the residuals is right skewed… condition not met.

Diagnostics

What do you think? *(Response Variable = **Income**)*

- **Independence of Residuals Condition**:
  - No clear structure/pattern in residuals as order of data collection ("time") increases, so this data has no time-series structure… so independence of residuals not violated via a time series relationship.
  - Should also check to see if data is collected via random sampling /assignment and meets the 10% rule, if not this would also violate independence.

4



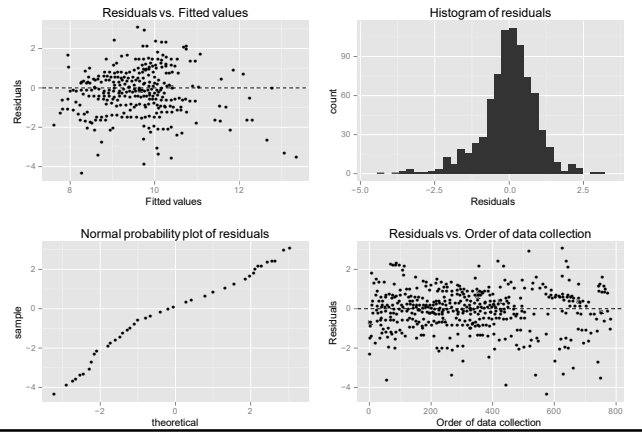
Diagnostics -- code

```
m_full_log = lm(log(income) ~ hrs_work + race + age + gender
    + citizen, data = acs_emp)
```

```
# residuals vs. fitted
qplot(data = m_full_log, y = .resid, x = .fitted, geom = "point") +
  geom_hline(yintercept = 0, linetype = "dashed") +  xlab("Fitted values")
  +
  ylab("Residuals") +
  ggtitle("Residuals vs. Fitted values")

# histogram of residuals
qplot(data = m_full_log, x = .resid, geom = "histogram") +
  xlab("Residuals") +
  ggtitle("Histogram of residuals")

# normal prob plot of residuals
qplot(data = m_full_log, sample = .resid, stat = "qq") +
  ggtitle("Normal probability plot of residuals")

# order of residuals
qplot(data = m_full_log, y = .resid) +  geom_hline(yintercept =
  0, linetype = "dashed") +  ylab("Residuals") +
  xlab("Order of data collection") +
  ggtitle("Residuals vs. Order of data collection")
```

Log transformation

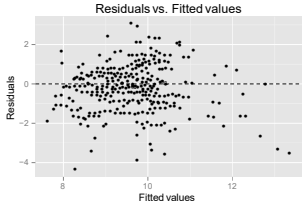
What do you think? *(Response Variable = **log(Income)**)*



Diagnostics

What do you think? *(Response Variable = **log(Income)**)*

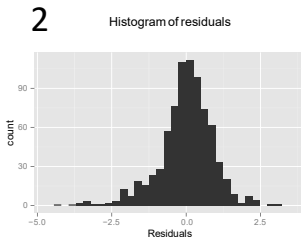
1

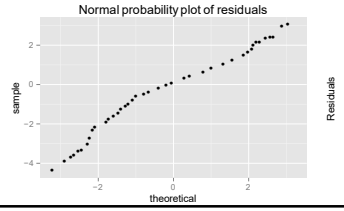


- **Linearity Condition**: Scatter distributed roughly evenly above/below the line as fitted values increase. → **roughly linear**.
- **Constant Variance of Residuals Condition**: Variance of residuals roughly constant as fitted values increases → **constant variance is met.**

Diagnostics

What do you think? *(Response Variable = **log(Income)**)*



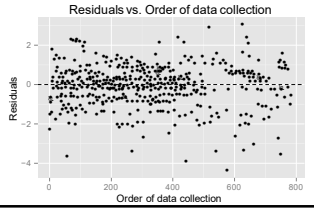


- **Normality of Residuals Condition**: The distribution of the residuals is roughly normal and centered at 0. Condition met.

Diagnostics

What do you think? *(Response Variable = **log(Income)**)*

- **Independence of Residuals Condition**:
  - No clear structure/pattern in residuals as order of data collection ("time") increases, so this data has no time-series structure… so independence of residuals not violated via a time series relationship.
  - Should also check to see if data is collected via random sampling /assignment and meets the 10% rule, if not this would also violate independence.



Application exercise: 7.2 Interpreting models with a transformed response

See course website for more details

		Iteration 0 (Full Model)	Iteration 1	Iteration 2	
	Highest Adjusted R^2 from the List of Actions	0.4388919	0.4390839	0.4339406	
	What to do next?	**Take note of Adjusted R^2 and move on to Iteration 1.**	**0.4390839 is higher than 0.4388919 … so PERMANENTLY REMOVE CITIZEN from the model and move on to Iteration 2.**	**0.4339406 is NOT higher than 0.4388919 … so DON'T REMOVE RACE and stick with the model you had in the previous iteration. End the algorithm.**	
Sub-Iteration Actions	Action: Remove hrs_work Variable from Model and Rerun		Adjusted R^2 =0.1700819	Adjusted R^2 =0.1707284	**STOP ALGORITHM!**
	Action: Remove race Variable from Model and Rerun		Adjusted R^2 =0.4339637	**Adjusted R^2 =0.4339406**	
	Action: Remove age Variable from Model and Rerun		Adjusted R^2 =0.3746502	Adjusted R^2 =0.3754503	
	Action: Remove gender Variable from Model and Rerun		Adjusted R^2 =0.4303261	Adjusted R^2 =0.430434	
	Action: Remove citizen Variable from Model and Rerun		**Adjusted R^2 =0.4390839**		